wCorr Formulas

Paul Bailey, Ahmad Emad, (people who do QC for this product)

2016-03-09

This document shows the methodology used in computing the correlations in wCorr for the weighted Pearson, Spearman, polyserial, and polychoric correlations. For the polyserial and polychoric correlations the coefficient is estimated using a numerical likelihood maximization. The weighted likelihood functions that are used are motivated and described.

Additionally, for the polyserial and polychoric evidence, is offered for the correctness of the methods, including a sketch of a proof of the consistency of both methods and numerical simulations to show:

- The bias of the methods as a function of the true correlation coefficient (ρ) and the number of observations (n)
- The root mean square error (RMSE) of the methods as a function of ρ and n in the weighted and unweighed cases

The *wCorr Arguments* vignette describes the effects the `ML` and `fast` arguments have on computation.

Methodology

Here we focus on measurement of the correlation coefficients between two vectors of random variables that are jointly bivariate normal–call the vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$.[1] The i^{th} members of the vectors are then called x_i and y_i.

Pearson and Spearman methodology

The Pearson correlation is computed using the formula

$$\rho_{Pearson} = \frac{\sum_i [w_i(x_i - \bar{x}) \times (y_i - \bar{y})]}{\sqrt{\sum_i (w_i \times (x - \bar{x})^2) + \sum_i (w_i \times (y - \bar{y})^2)}}$$

where $\bar{x} = \frac{1}{\sum w_i} \sum_i w_i x_i$, $\bar{y} = \frac{1}{\sum w_i} \sum_i w_i y_i$, and n is the number of elements in $\boldsymbol{X}$ and $\boldsymbol{Y}$.[2]

The Spearman correlation coefficient is calculated by first taking the rank of the data before the same formula is applied. When data are ranked ties must be handled in some way. The chosen method is to assign the average of all tied ranks. For example, if the second and third rank units are tied then both units would receive a rank of 2.5 (the average of 2 and 3).

Polyserial methodology

For the polyserial correlation, it is again assumed that there are two continuous variables $\boldsymbol{X}$ and $\boldsymbol{Y}$ that have a bivariate normal distribution.[3]

[1] The Spearman correlation coefficient can be motivated by a much more broad class of random variables but that is not dealt with in this document.

[2] See the "correlate" function in Stata Corp, Stata Statistical Software: Release 8. College Station, TX: Stata Corp LP, 2003.

[3] For a more complete treatment of the polyserial correlation, see Cox, N. R., "Estimation of the Correlation between a Continuous and a Discrete Variable" *Biometrics*, **50** (March), 171-187, 1974.

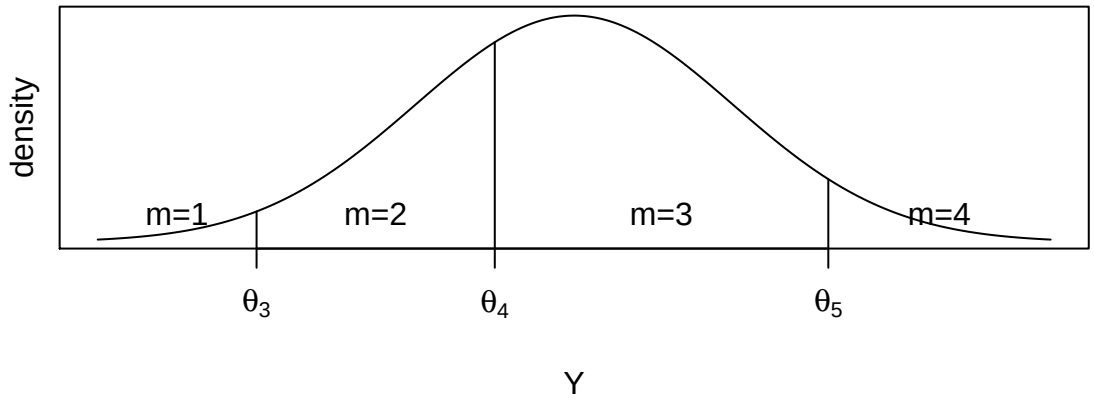
$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim N\left[\begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}, \Sigma\right]$$

Where $N(a, \Sigma)$ is a bivariate normal distribution with mean a and variance Σ. For the polyserial, $\boldsymbol{Y}$ is discretized into the random variable $\boldsymbol{M}$ according to

$$m_i = \begin{cases} 1 & \text{if} \theta_2 < y_i < \theta_3 \\ 2 & \text{if} \theta_3 < y_i < \theta_4 \\ \vdots & \\ t & \text{if} \theta_{t+1} < y_i < \theta_{t+2} \end{cases}$$

where, for notational convenience, $\theta_2 \equiv -\infty$ and $\theta_{t+2} \equiv \infty$.[4]

The following figure shows the density of $\boldsymbol{Y}$ when $\theta = (-\infty, -2, -0.5, 1.6, \infty)$. Here, for example, any value of $-2 < y_i < -0.5$ would have $m_i = 2$.



Notice that μ_y is not identifed (or is irrelevant) because setting $\tilde{\mu}_y = \mu_y + a$ and $\tilde{\boldsymbol{\theta}} = \boldsymbol{\theta} + a$ lead to exactly the same values of $\mathbf{M}$ and so one of the two must be artibrary assigned. A convenient decision is to decide $\mu_y \equiv 0$. A similar argument holds for σ_y so that $\sigma_y \equiv 1$.

For $\boldsymbol{X}$ Cox (1974) observes that the MLE mean and standard deviation of $\boldsymbol{X}$ are simply the average and (population) standard deviation of the data and do not depend on the other parameters.[5] This can be taken advantage of by defining $z \equiv \frac{x - \bar{x}}{\hat{\sigma}_x}$.

Combining these simplifications the probability of any given x_i, m_i pair is

$$\Pr\left(\rho = r; Z = z_i, M = m_i\right) = \phi(z_i) \int_{\theta_{m_i+1}}^{\theta_{m_i+2}} dy f(y|Z = z, \rho = r)$$

where $\Pr\left(\rho = r; Z = z, M = m\right)$ is the probability of the event $\rho = r$ given the data z and m, $\phi(\cdot)$ is the standard normal and $f(y|z, \rho)$ is the distribution of y conditional on z and rho. Because y and z are jointly

[4]The indexing is somewhat odd to be consistent with Cox (1974). Nevertheless, this treatment does not use the Cox definition of θ_0, θ_1 or θ_2 which are either not estimated (as is the case for θ_0, and θ_1) or are reappopriated (as is the case for θ_2). Cox calls the correlation coefficient θ_2 while this document uses ρ and uses θ_2 to store $-\infty$ as a convenience so that the vector $\boldsymbol{\theta}$ includes the (infinite) bounds as well as the interior points.

[5]The population standard devaition is used because it is the MLE for the standard deviation. Notice that, while the sample variance is an unbiased estimator of the variance and the population variance is not an unbaised estimator of the variance, they are very similar and the variance is also a nusiance parameter, not a parameter of interest when finding the correlation.

bivariate normally distributed (by assumption)

$$f(y|Z=z,\rho=r)=N\left(\mu_y+\frac{\sigma_y}{\sigma_z}r(z-\mu_z),(1-r^2)\sigma_y\right)$$

because both $\boldsymbol{Z}$ and $\boldsymbol{Y}$ are standard normals

$$f(y|Z=z,\rho=r)=N\left(r\cdot z,(1-r^2)\right)$$

now, define $w\equiv\frac{y-r\cdot z}{1-r^2}$ and w has a standard normal distribution. Plugging this in

$$\Pr\left(\rho=r,\theta;Z=z_i,M=m_i\right)=\phi(z_i)\left[\Phi\left(\frac{\theta_{m_i+2}-r\cdot z_i}{\sqrt{1-r^2}}\right)-\Phi\left(\frac{\theta_{m_i+1}-r\cdot z_i}{\sqrt{1-r^2}}\right)\right]$$

Where $\Phi(\cdot)$ is the standard normal cumulative density function. Using the above probability function as an objective, the log-likelihood is then maximized.

$$\ell(\rho;z,m)=\sum_i w_i\ln\left[\Pr\left(\rho=r;Z=z_i,M=m_i\right)\right]$$

where w_i is the weight of the i^{th} case.

The value of the nusiance paramter $\boldsymbol{\theta}$ is chosen to be $\Phi^{-1}(n/N)$ where n is the number of values to the left of the cut point (θ_i value) and N is the number of data points overall.

Polyserial computation

The derivatives of ℓ can be computed but are not readily computed and so when the `ML` argumet is set to `FALSE` (the default) a one dimensional optimization of ρ is calculated using `stats::optimize`. When the `ML` argument is set to `TRUE` a multi-dimensional optimization is done for ρ and $\boldsymbol{\theta}$ using `minqa::bobyqa`. As is shown below, the difference between these two is slight, if present, and so the default value of `ML` is recommended.

Because the optimization is not perfect when the correlation is in a boundary condition ($\rho\in\{-1,1\}$), a check for perfect correlation is performed before the above optimization by simply seeing if the values of $\boldsymbol{X}$ and $\boldsymbol{M}$ have exactly the same order.

Polychoric methodology

Similar to the polyserial, the polychoric is a simple case of two continuous variables $\boldsymbol{X}$ and $\boldsymbol{Y}$ that have a bivariate normal distribution. In the case of the polyserial the continuous (latent) variable $\boldsymbol{Y}$ was observed as a discretized variable $\boldsymbol{M}$. For the polychoric this is again true but now the continuous (latent) variable $\boldsymbol{X}$ is observed as a discrete variable $\boldsymbol{P}$ according to

$$p_i=\begin{cases}1 & \text{if}\theta'_2<x_i<\theta'_3\\ 2 & \text{if}\theta'_3<x_i<\theta'_4\\ \vdots & \\ t & \text{if}\theta'_{t+1}<x_i<\theta'_{t+2}\end{cases}$$

where $\boldsymbol{\theta}$ remains the cut points for the distibution defining the transformation of $\boldsymbol{Y}$ to $\boldsymbol{M}$ and $\boldsymbol{\theta}'$ is the cut points for the transformation from $\boldsymbol{X}$ to $\boldsymbol{P}$. Similar to $\boldsymbol{\theta}$, $\boldsymbol{\theta}'$ has $\theta'_2\equiv-\infty$ and $\theta'_{t+2}\equiv\infty$.

Similar to in the polyserial, μ_y is not identifed (or is irrelevant) because setting $\tilde{\mu}_y=\mu_y+a$ and $\tilde{\boldsymbol{\theta}}=\boldsymbol{\theta}+a$ lead to exactly the same values of $\mathbf{M}$ and so one of the two must be artibrary assigned. The same is true

for μ_x. A convenient decision is to decide $\mu_y = \mu_x \equiv 0$. A similar argument holds for σ_y and σ_x so that $\sigma_y = \sigma_x \equiv 1$

Then the probability of any given m_i, p_i pair is

$$\Pr\left(\rho = r; P = p_i, M = m_i\right) = \int_{\theta'_{p_i+1}}^{\theta_{p_i+2}} dx \int_{\theta_{m_i+1}}^{\theta_{m_i+2}} dy f(x, y|\rho = r)$$

where ρ is the correlation coefficient.

Using this function as an objective, the log-likelihood is then maximized.

$$\ell(\rho; p, m) = \sum_i w_i \ln\left[\Pr\left(\rho = r; P = p_i, M = m_i\right)\right]$$

Polychoric computation

This again mirrors the treatment of the polyserial. The derivatives of ℓ can be computed but are not readily computed and so when the `ML` argument is set to `FALSE` (the default) a one dimensional optimization of ρ is calculated using `stats::optimize`. When the `ML` argument is set to `TRUE` a multi-dimensional optimization is done for ρ, $\boldsymbol{\theta}$, and $\boldsymbol{\theta}'$ using `minqa::bobyqa`. As is shown below, the difference between these two is slight, if present, and so the default value of `ML` is recommended.

Because the optimization is not perfect when the correlation is in a boundary condition ($\rho \in \{-1, 1\}$), a check for perfect correlation is performed before the above optimization by simply seeing if the values of $\boldsymbol{P}$ and $\boldsymbol{M}$ have a Goodman-Kruskal correlation coefficient of -1 or 1. When this is the case, the MLE of -1 or 1, respectively, is returned.

Correctness

It is easy to prove the consistency of the $\boldsymbol{\theta}$ for the polyserial and $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$ using the non-ML case. Similarly, for ρ, because it is an MLE that can be obtained by taking a derivative and setting it equal to zero, the results are asymtotically unbiased and obtain the Cramer-Rao lower bound.

This does not speak to the small sample properties of these correlation coefficients. Previous work has described their properties by simulation and so that tradition is continued below.

General setup for the unweighted case

A simulation is run several times. For each itteration, the following procedure is used:

- select the number of observations (n)
- select a true correlation coefficient ρ
- generate $\boldsymbol{X}$ and $\boldsymbol{Y}$ to be bivariate normally distributed using a pseudo-Random Number Generator (RNG)
- using a pseudo-RNG, select the the number of bins for $\boldsymbol{M}$ and $\boldsymbol{P}$ (t and t') independantly from the set {2, 3, 4, 5}
- select the bin boundaries for $\boldsymbol{M}$ and $\boldsymbol{P}$ ($\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$) by sorting the results of $(t-1)$ and $(t'-1)$ draws, respectively, from a normal distribution using a pseudo-RNG
- confirm that at least 2 levels of each of $\boldsymbol{M}$ and $\boldsymbol{P}$ are occupied (if not, retrun to generating $\boldsymbol{X}$ and $\boldsymbol{Y}$)
- calculate and record relevant statistics

When the exact method of selecting a parameter (such as n) is not noted in the above description it is described as part of each simulation.

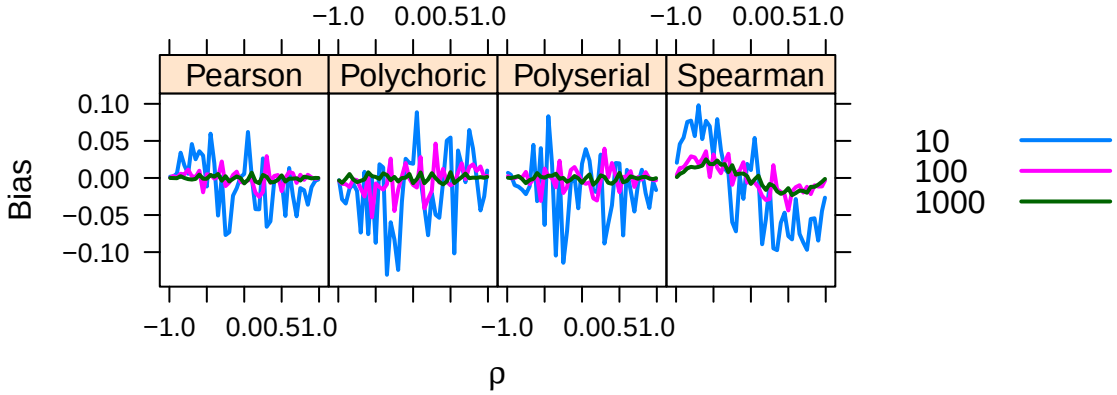
bias of the correlations

This sections shows the bias of the correlations as a function of the true correlation coefficient, ρ. To that end, a simulation was done at each level of the cartesian product of $\rho \in (-0.99, -0.95, -0.90, -0.85, ..., 0.95, 0.99)$, and $n \in \{10, 100, 1000\}$. For precision, each iteration is run fifty times. The bias is the mean difference between the true correlation coefficient (ρ_i) and estimate correlation coefficient (r). the RMSE is the squareroot of the mean square error.

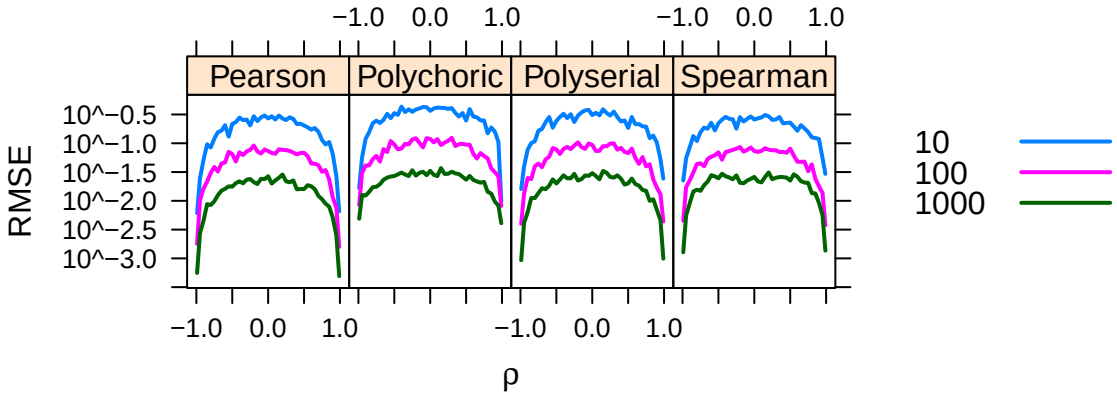
$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_i (r_i - \rho_i)^2}$$

RMSE of the correlations as a function of n and ρ

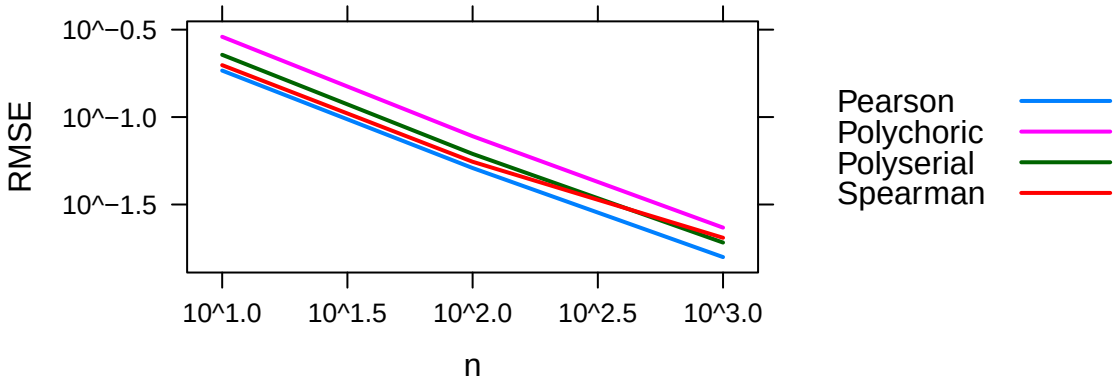
This plot shows the bias as a function of the true correlation ρ. Only the Spearman correlation shows a clear trend with a positive bias below 0 and a negative bias above zero.



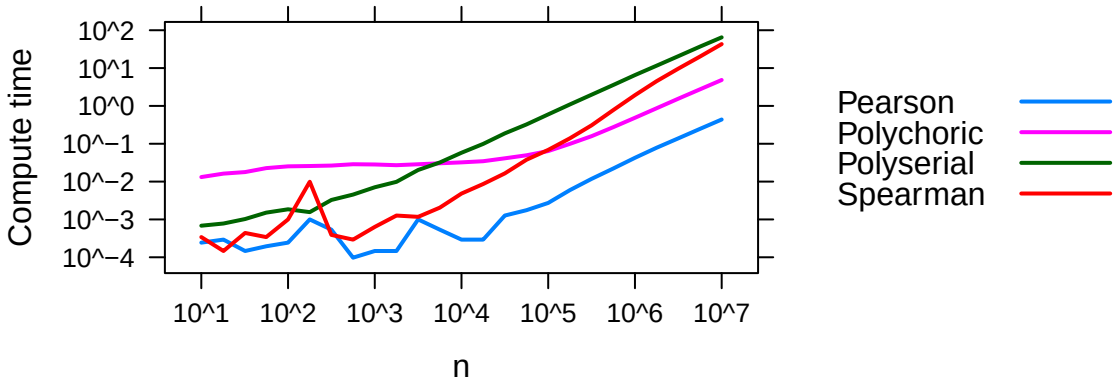
this plot shows the RMSE as a function of ρ. All of the correlation coefficients have a uniform RMSE as a function of ρ near $\rho = 0$ that decreases near $|\rho| = 1$. All plots also show a decrease in RMSE as n increases. That is further shown in the next plot.



This plot shows the RMSE as a function of n. The plot shows a slope of about $-\frac{1}{2}$, which suggests first order consistency.



Finally, this plot shows the mean time to compute a single correlation coefficient as a function of ρ by n size.



General setup for the weighted case

In this section the consistency of the weighted coefficients is explored.

For the weighted case the unweighted result would be consistent if there were not something about the higher weight cases that makes them diffeent from the lower weight cases. Thus, while it is not reasonable to always assume that thee is a difference between the high and low weight cases, boht are possible and it serves as a more robust test of the methods in this package to consider a cases where they are associated. So, when sampling (other than simple random sampling with replacement) is used unweighted correlations may be consistent or may not be. The simulations below expore a case where the unweighted are not expected to be consistent to see if the weighted results are consistent.

Weighted simulation study

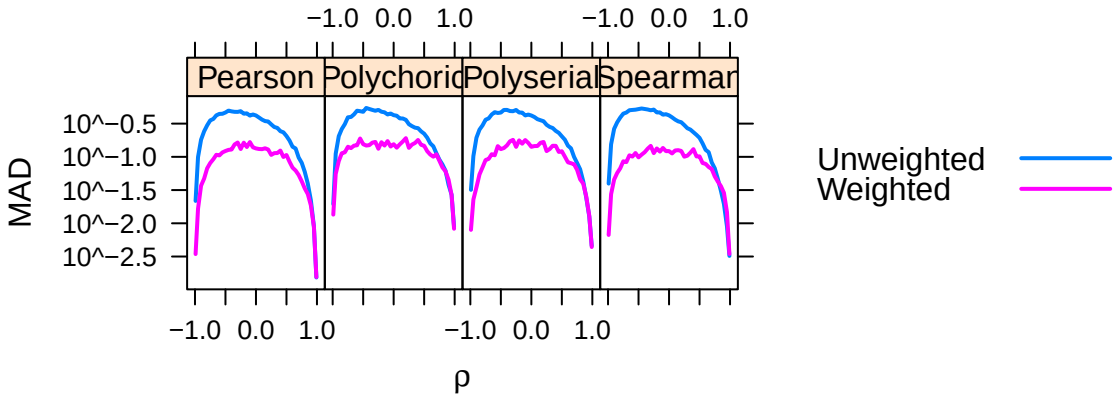
Simulations are carred out in the same fashion as before with a few extra steps to accomodate weights. The following changes were made:

- to find the correlation between n values, $5n$ values were generated

- weights are assigned according to $w_i = (x - y)^2 + 1$ and the probability of inclusion in the sample was proportional to $Pr_i = \frac{\sum_j w_j}{w_i}$.
- a sample of size n was then selected from the $5n$ values according to the proabilities and weighted as well as unweighted correlation coefficients were calculated.

Two simulations were run. The first shows the MAD as a function of ρ and was run for $n = 100$ and $\rho \in (-0.99, -0.95, -0.90, -0.85, ..., 0.95, 0.99)$, with 100 iterations run for each value of ρ.

The following plot shows the MAD for the weighted and unweighted results as a function of ρ when $n = 100$.



The second simulation used the same values of ρ and used $n \in \{10, 100, 1000, 10000, 100000\}$ and shows how RMSD and sample size are related. In particular, it shows first order convergence of the weighted correlation coefficient. Again, the fact that the simulations show that the unweighted correlation coefficient is not consistent is not meant to imply that it will always be that way–only that this is possible.